A Feature Map Adversarial Attack Against Vision Transformers

Majed Altoub[1], Rashid Mehmood[2], Fahad AlQurashi[3], Saad Alqahtany[4], Bassma Alsulami[5]
Department of Computer Science, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah 21589, Saudi Arabia[1,3,5]
Department of Computer Science, Faculty of Computer and Information Systems
Islamic University of Madinah, Madinah 42351, Saudi Arabia[2,4]

***Abstract*—Image classification is a domain where Deep Neural Networks (DNNs) have demonstrated remarkable achievements. Recently, Vision Transformers (ViTs) have shown potential in handling large-scale image classification challenges by efficiently scaling to higher resolutions and accommodating larger input sizes compared to traditional Convolutional Neural Networks (CNNs). However, in the context of adversarial attacks, ViTs are still considered vulnerable. Feature maps serve as the foundation for representing and extracting meaningful information from images. While CNNs excel at capturing local features and spatial relationships, ViTs are better at understanding global context and long-range dependencies. This paper proposes a feature map ViT-specific adversarial example attack called Feature Map ViT-specific Attack (FMViTA). The objective of the investigation is to generate adversarial perturbations in the spatial and frequency domains of the image representation that allow deeper distance measurement between perturbed and targeted images. The experiments focus on a ViT pre-trained model that is fine-tuned on the ImageNet dataset. The proposed attack demonstrates the vulnerability of ViTs to adversarial examples by showing that even allowing only 0.02 maximum perturbation magnitude to be added to the input samples gives 100% attack success rate.**

Keywords*—Vision transformers; adversarial attacks; DNNs; vulnerabilities; feature maps; perturbations; spatial domains; frequency domains

I. Introduction

Deep Neural Networks (DNNs) have emerged as highly effective image classification tools. Convolutional Neural Networks (CNNs) recognize the input image data by the convolutional layers to identify and capture local patterns and spatial hierarchies of features [1]. On the other hand, Vision Transformers (ViTs) represent the input image data as sequences of patches and leverage the self-attention mechanisms to capture long-range dependencies and global context in images [2]. Therefore, ViTs can be more effective than CNNs in long-range interactions and achieve state-of-the-art performance on various image classification benchmarks.

Feature maps are fundamental for representing and extracting valuable information from images in DNN classification models. ViTs are better at understanding global context and long-range dependencies than CNNs [3]. ViTs divide the input image into patches and then flatten and embed them into tensors. These tensors are then fed into a transformer encoder, which learns to attend to different patches and their relationships.

From the security perspective, the adversarial robustness of the ViT models is a significant challenge, specifically in a domain such as image classification. Adversarial attacks intentionally manipulate the models by inserting small perturbations into the input image data to deceive them, resulting in incorrect predictions of the input image while appearing normal to humans, called adversarial example attacks. Adversarial perturbations are critical in understanding the vulnerabilities of the ViT models to adversarial examples.

The existent studies of adversarial perturbations have focused on two main areas: the spatial domain and the frequency domain [4]. In the spatial domain, the representation of the input images is extracted directly from the pixel values which is typically used in attacks such as gradient-based adversarial attacks [5]–[10]. The frequency domain including low-frequency and high-frequency is another perspective that can be used to generate perturbations in selected frequency regions [4]. The representation of the input images in the frequency domain is transformed through methods like the Discrete Cosine Transform (DCT) [11]. Prior research has indicated that CNNs are vulnerable to high-frequency noise and ViTs are vulnerable to low-frequency noise [12]. These findings have contributed to the emergence of frequency-based adversarial attacks, including [11], [13]–[17].

In this research paper, our objective is to investigate the potential impact of introducing an adversarial example attack by generating the adversarial perturbations in the spatial domain but with some potential influence from the frequency domain. We introduce the FMViTA attack that uses the ViT feature maps of the perturbed and the target images to optimize the adversarial perturbations. For deeper comparison, the cosine similarity loss function is used to measure the similarity between feature maps. The ViT model itself can learn to capture frequency-related information through its self-attention mechanism. This means that the feature maps used for cosine similarity might contain some frequency-based representations, especially since we take the mean feature maps from the intermediate blocks. While the ViT feature maps are not considered directly spatial domains, they are still influenced by the spatial arrangement of the image patches. However, adding adversarial perturbations to the input image is a direct pixel-level operation happening in the spatial domain. Furthermore, for imperceptibly added, we clamp adversarial perturbations.

We, however, do not intend to make a strength transferable adversarial example attack; we pose a direction of generating perturbation by using more deeper comparison methods to optimize the perturbations generated. Our results demonstrate

that using FMViTA method can indeed manipulate a ViT model to classify a set of five input images as other target images with 100% attack success rate, while we only allowing 0.02 maximum perturbation magnitude to be changed. The contributions of this paper are summarized as follows:

- We employ a cosine similarity loss function for deeply measuring the similarity between the perturbated and the target feature maps of the images to optimize the perturbation patterns.
- We also propose a novel Feature Map ViT-specific Attack method named FMViTA that takes one input image and runs it against a target image to generate a perturbated image that looks like the input one but classifies as the target image.
- We demonstrate the results of using FMViTA against a ViT for five input and target images.

The rest of the paper is organized as follows: Section II briefly gives a background and reviews the related works. In Section III, we provide our methodology, including the threat model. Experiments setup and results are presented in Section IV. Finally, the conclusion is made in Section V.

II. Background and Related Work

A. Adversarial Attacks in DNNs

The adversarial attacks can be broadly classified by the attack goals to evasion, poisoning, backdoor, and privacy attacks. In evasion attacks, the goal is to cause misclassification at test time, such as in the adversarial example attacks [18]. In the poisoning attacks, the goal is to inject poisonous data that can manipulate the DNN model's training. The goal of the privacy attacks is to steal information from the DNN models. Backdoor attacks can be considered both evasion and poisoning attacks because adversaries inject poisonous data called triggers into legitimate DNN models at training time and misclassify the model once the trigger is activated at test time [19].

From another point of view, adversarial attacks can be classified based on the accessibility of the adversaries. In this context, adversarial attacks can be either white-box or black-box attacks. White-box attacks are powerful because adversaries can access the architecture and parameters of a DNN model. On the other hand, black-box attacks are more challenging because the adversaries can only access the input and output data from a DNN model. Gray-box attacks are situated between white-box and black-box attacks. In these attacks, the adversaries possess limited knowledge of the model's architecture and parameters.

In addition, depending on the goals of the attacks, adversarial attacks can be classified as targeted and untargeted. In untargeted attacks, the goal is to misclassify the DNN classification model to output any other class. In the targeted attacks, the goal is to misclassify the DNN classification model to classify the input as a particular class. In this paper, our attack method is considered a targeted gray-box adversarial example attack. However, the method could be developed to generate other types of attacks.

B. Adversarial Example Attack Methods

The earlier adversarial example attacking phenomena were proposed by Szegedy et al. [18]. Since then, various methods for creating highly powerful perturbations have been put forward in academic literature. Goodfellow et al. [5] first introduced the concept of adversarial examples using a gradient-based method called the Fast Gradient Sign Method (FGSM). In this method, the gradient of the loss function is used with respect to the input data to generate the small perturbation that maximizes the loss function, which can be conceded as a white-box attack. On the other hand, Kurakin et al. [6] improve the FGSM by running multiple iterations of the attack. However, the method showed that there would still be a gap in accuracy between the perturbed results and the being ones. Later researchers use momentum [7] approaches to further enhance the adversarial example attacks. Dong et al. [8] proposed the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) to increase the success rate of the black-box attacks, which improves the transferability of the adversarial example attacks. Other enhancements were proposed as well, such as the Ensemble Momentum Iterative FGSM (EMI-FGSM) [9] and VMI-FGSM and VNI-FGSM [10].

Frequency-based adversarial example attacks have been developed in the literature to generate adversarial perturbations. Recently, Duan et al. [11] proposed another way of crafting adversarial examples named AdvDrop. Instead of adding perturbations, they drop existing details from clean images from frequency components. The results demonstrate that AdvDrop can achieve high attack success rates on ImageNet dataset. Guo et al. [15] focus on the low frequency adversarial perturbation for black-box attacks. The authors demonstrate the effectiveness of this technique by successfully fooling the Google Cloud Vision platform with an unprecedented low number of only 1000 model queries . Jin et al. [17] introduced a novel white-box attack known as the Frequency and Spatial Consistency Based Adversarial Attack (FSA). The results show that the FSA method can enhance the success attack rates, with a maximum improvement of 28.98% observed across various attack methods and models.

C. Adversarial Robustness of Vision Transformer

ViT is state-of-the-art for image classification that uses a transformer architecture to process image patches and generate predictions [20]. As ViTs continue to gain popularity and are being deployed in various domains, the robustness of ViTs to adversarial attacks becomes essential.

According to Aldahdooh et al. [21] Vanilla ViTs or hybrid ViTs are more robust to adversarial attacks than CNNs, particularly under Lp-norm or adaptive attacks. The findings demonstrate that increasing the number of attention blocks may increase the robustness to transfer attacks but not white-box attacks. Shao et al. [22] claim that ViTs are less sensitive to high-frequency perturbations than CNNs and MLP-Mixers. That is because ViTs contain less high-frequency features, which have a high correlation with their robustness against different frequency-based perturbations. In addition, in ViTs, introducing convolutional or token-to-token blocks for learning high-frequency features can improve classification accuracy, but at the cost of adversarial robustness. Furthermore, modern

CNNs may borrow some of ViTs techniques to bridge the performance gap as well as the adversarial robustness gap.

Paul and Chen [23] have compared the robustness of ViTs and CNNs models against frequent corruptions and perturbations using six different ImageNet datasets. Their findings suggest that ViTs have better robustness performance than CNNs on some of the datasets, and that the robustness of both models can be improved by fine-tuning on corrupted data. Overall, the study highlights the potential of ViTs as a promising alternative to CNNs in computer vision tasks, especially when it comes to robustness. Joshi et al. [24] analyzed the underlying distinctions between the ViT and CNN models. Unlike convolutional networks, vision transformers use a patch token-based self-attention mechanism. By creating a block sparsity based adversarial token attack, their study found that ViT models are more sensitive to token attacks than CNN models. This highlights the need for robustness evaluation of transformer-based models against adversarial attacks. It also suggests that further research is required to improve the security of vision transformers.

III. Methodology

A. Threat Model

We use the threat model to define our FMViTA attack scenario. We focus on image classification; nevertheless, the technique can be easily extended to other domains. In this threat model, we assume that the adversary has limited knowledge of the model's architecture and parameters. In addition, the adversary does not have access to the training data. The attack is considered a gray-box adversarial attack. The adversary's objective in this threat model is to embed perturbations into the input images to make them perturbed in order to fool the ViT classification model to classify them as target images' classes.

B. Attack Method

We formulate FMViTA as an optimization problem to generate perturbed images. The optimization is based on minimizing the distance between feature representations of the perturbed and target images. To do so, we take the mean of the intermediate features for both input and target images. Then we calculate the cosine similarity between the mean features of the perturbed and target images to measure the distance between them, which is the loss function to be used in the gradient iterations.

The loss function encourages the features of the perturbed image to be similar to those of the target image. Our attack utilizes the intermediate feature maps instead of focusing on the model's output probabilities. Thus, we chose the cosine similarity as a function due to its suitability for comparing high-dimensional feature vectors. Furthermore, the cosine similarity concentrates on the angular relationship between the two feature maps. Compared to other similarity calculations, such as Euclidean and Manhattan distances, cosine similarity is less sensitive to magnitude differences because it focuses on the angle between vectors. This property allows for a more in-depth comparison between the two feature maps. Since the cosine similarity is used to measure this similarity, a higher similarity results in a lower loss as shown in Eq. (1).

$$L(F_t, F_p) = 1 - \frac{F_t \cdot F_p}{\|F_t\|\|F_p\|} \tag{1}$$

$L(F_t, F_p)$: The loss function measuring the difference between the feature representations of the target image F_t and the perturbed image F_p.

$F_t \cdot F_p$: The dot product of the feature vectors, representing their similarity.

$\|F_t\|$and$\|F_p\|$: The magnitudes (norms) of the feature vectors, normalizing the cosine similarity.

Let M be the target ViT model and I_t and I_s be the input and target images. The optimizing perturbed patterns P is added to the input image as to generate perturbed images I_p as shown in Eq. (2). In addition, we clamp the perturbed pattern values to prevent them from becoming too large as shown in Eq. (3).

The gradient optimization can be mathematically described as:

$$\nabla_{t_i} L_i = \nabla_{t_i} \left(1 - L\left(mean\left(F_p\left(I_p\right)\right), mean(F_t(I_t))\right)\right) \tag{2}$$

$$I_p = \text{clamp}(I_s + P, -0.02, 0.02) \tag{3}$$

Algorithm 1: Feature Map ViT-specific Attack (FMViTA)

Input: Target model M, target image I_t, input images I_s, perturbed pattern P. **Output:** Perturbed images I_p.

Initialize P zeros.

for *each epoch* e **do**

  for *each input image* I_s *in* I_s **do**

    $I_p \leftarrow \text{Clamp}(I_s + P)$ Generate perturbed image

    $F_t \leftarrow M(I_t)$ Get features of target image

    $F_p \leftarrow M(I_p)$ Get features of perturbed image

    $L \leftarrow \text{Loss}(F_t, F_p)$ Calculate loss

    Update P using gradient descent: $P \leftarrow P - \alpha \nabla L$

  Return: Perturbed images I_p

IV. Experiments and Results

In this section, we first explain our experimental setup and then show the experimental results.

A. Experimental Setup

Our experiments mainly focus on pre-trained ViT (google/vit-base-patch16-224) [25], [26] that was pre-trained on ImageNet-21k [27] and fine-tuned on ImageNet [26] with a resolution of 224x224 and 1000 classes. Instead of using Adam, we use AdamW optimizer with a learning rate of 0.06 and weight decay of 0.002. The experiments were conducted on a dataset consisting of various AI-generated images for the input and target images that are classified as the ImageNet labels. We use AdamW optimizer to add the regularization

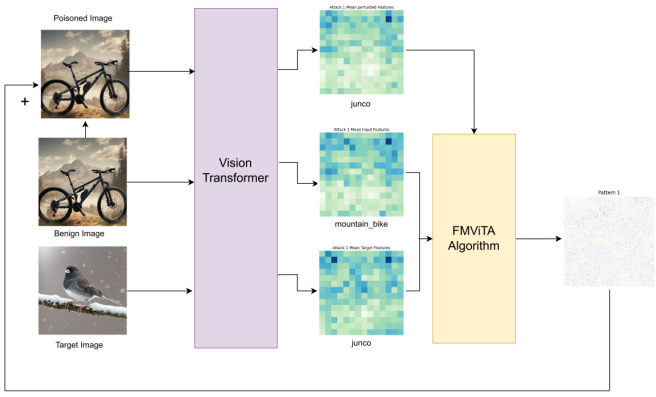

Fig. 1. The overall view of FMViTA attack ruining for Attack 1.

benefits of L2 regularization. We use AI-generated images as a dataset samples to be the target, and for the input images, we use five classes from the ImageNet dataset, which are {junco, mountain bike, terrapin, soccer ball, and goldfinch}.

Adam is an algorithm for first-order gradient-based optimization that computes adaptive learning rates for different parameters by maintaining estimates of the first and second moments of the gradients [28]. Adam combines the advantages of AdaGrad and RMSProp, with key differences in how it updates parameters and handles bias correction. However, L2 regularization and weight decay are equivalent for standard stochastic gradient descent (SGD) [29], but not for adaptive gradient algorithms like Adam. AdamW [30] modifies the popular Adam optimizer to improve its generalization performance, allowing it to compete with SGD with momentum on image classification tasks. To take advantage of AdamW we use it for the gradient-based optimization of FMViTA attack.

B. Experimental Results

We run the experiments to test the effectiveness of our attack algorithm using the same adjusted hyperparameters, such as the learning rate of 0.06, weight decay of 0.002, and 700 epochs for all five attacks. These hyperparameters were determined through an iterative process of empirical evaluation. The choice of a 0.06 learning rate suggests a balance between speed and stability for all five attacks to be run as a one-time attack; however, each individual attack can be run using a different learning rate based on the images being used. The weight decay parameter incorporated in the AdamW optimizer offers L2 regularization to address overfitting, consequently improving the generalization of the generated adversarial examples. It was set at 0.002 to provide the best balance between preventing overfitting and maintaining sufficient subtle adversarial perturbations. We decided to use 700 epochs after monitoring the convergence behaviors of the attack's loss functions.

In order to assess the effectiveness of the proposed FMViTA attack, we conducted an evaluation of the attack's capability to manipulate the source of five different input images and five corresponding target images. This paired selection allowed a more robust evaluation, and to determine how well the attack generalizes to unseen data. In Attack 1, we use a mountain bike as an input image, then after extracting the feature maps from the 12 blocks of the ViT model, we take the mean of all the maps to be in one tensor, and we do the same for the target image, which is a junco. Then we run the attack algorithm using cosine similarity as the loss function between the mean feature map of the perturbed image (the input image + perturbed pattern) and the mean feature map of the target image. We optimize the perturbed pattern using the AdamW optimizer to add the regularization benefits of L2. The goal is to make the mean feature map of the perturbed image as close as the mean feature map of the target image. The overall illustration of Attack 1 can be seen in the Fig. 1.

In Attack 2, we do it the opposite way, using a "junco" as an input image and a "mountain bike" as the target image. Then, we do the same as in Attack 1. In Attack 3 the input image was "soccer ball" and the target image was a "terrapin". In Attacks 4 and 5 "goldfinch" and "terrapin" were the input images, and "soccer ball" and "goldfinch" were the target images. All attacks' results are shown in Fig. 3.

To evaluate the attack performance, we use the attack success rate (ASR) to measure the success of the FMViTA attack. The ASR refers to the percentage of FMViTA that effectively induces a misclassification in the ViT model, and it can be represented as shown in Eq. (4).

$$ASR_{\text{FMViTA}} = \frac{N_{\text{mis}}}{N_{\text{t}}} * 100 \tag{4}$$

where: N_{mis} is the number of perturbed images from FMViTA that are misclassified by the ViT model. N_{t} is the total number of experiments.

In order to enhance the existing evaluation method, we have fixed the parameters for a group of five input images against the other five target images. Our results show that the attack achieved a success rate of 100%. However, in terms of the invisibility of the noise in perturbed images, the attack result can be better when using only one input image against one target image and adjusting parameters based on that.

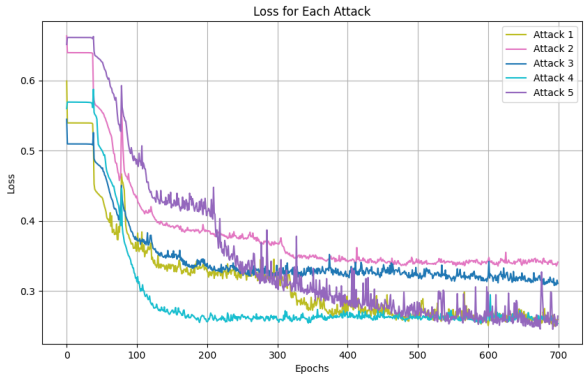

Fig. 2. The loss function results of the 5 attacks with 700 iterations.

That can be seen in Fig. 2 were losses behaved differently in each single attack iteration. Also, we found that extending the epochs did not improve the accuracy of in terms of the

invisibility of the noise in perturbed images, as if we changed the other parameters. The FMViTA attack tool will be available online with the flexibility to change the attack parameters.

Fig. 3. The results of FMViTA for a group of 5 input images against the other 5 target images.

Adversarial patterns often used in adversarial attacks, are specific inputs designed to elicit certain responses from a model. These patterns can expose the internal workings of neural networks and provide insights into their behavior under various conditions. Our patterns are typically generated through the FMViTA method to manipulate the input images intentionally.That can be seen in Fig. 4.

FMViTA method relies on feature maps, but the process of extracting features is computationally intensive, especially for long epochs and iterations in such an attack. Our optimization strategies are crucial for making the attack feasible within a reasonable timeframe. In FMViTA we extract the mean feature maps of the input and target images only once. However, it is necessary to extract the mean feature map of the perturbed images in each iteration. Fig. 5 shows the differences in mean feature maps of the input, target, and perturbed images.

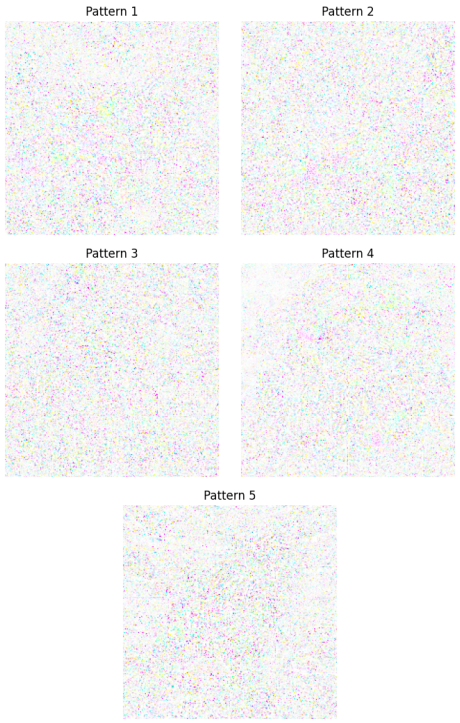

Fig. 4. The attack patterns for Attacks 1-5.

The vulnerability of ViTs to FMViTA stems from the exploitation of the intermediate feature representations within the DNN. The attack's efficacy derives from targeted manipulation of the intermediate features of the input image, leveraging gradient-based optimization, to generate perturbations that maximize the similarity between these intermediate features and the corresponding intermediate features of the target image. Thus, the success of FMViTA highlights the susceptibility of ViTs to gradient-based attacks targeting intermediate representations. Therefore, robust architectures and defenses against such sophisticated manipulation techniques are needed.

Mitigating the effectiveness of this attack approach suggests focusing on feature-space defenses. That involves analyzing the intermediate feature representations within ViTs to detect or mitigate the adversarial perturbations. One possible strategy is to apply denoising techniques to the intermediate feature maps. Recently, Yang et al. [31] have proposed the Denoising Vision Transformers (DVT) to suppress the grid-like artifacts observed in the feature maps by separating the clean features from those contaminated and then train a lightweight transformer block to predict clean features from raw ViT outputs. Another approach to mitigate such an attack is modifying the attention mechanism itself to enhance the robustness of attention weights to adversarial influence. That involves incorporating regularization terms during training or employing attention-aware denoising.

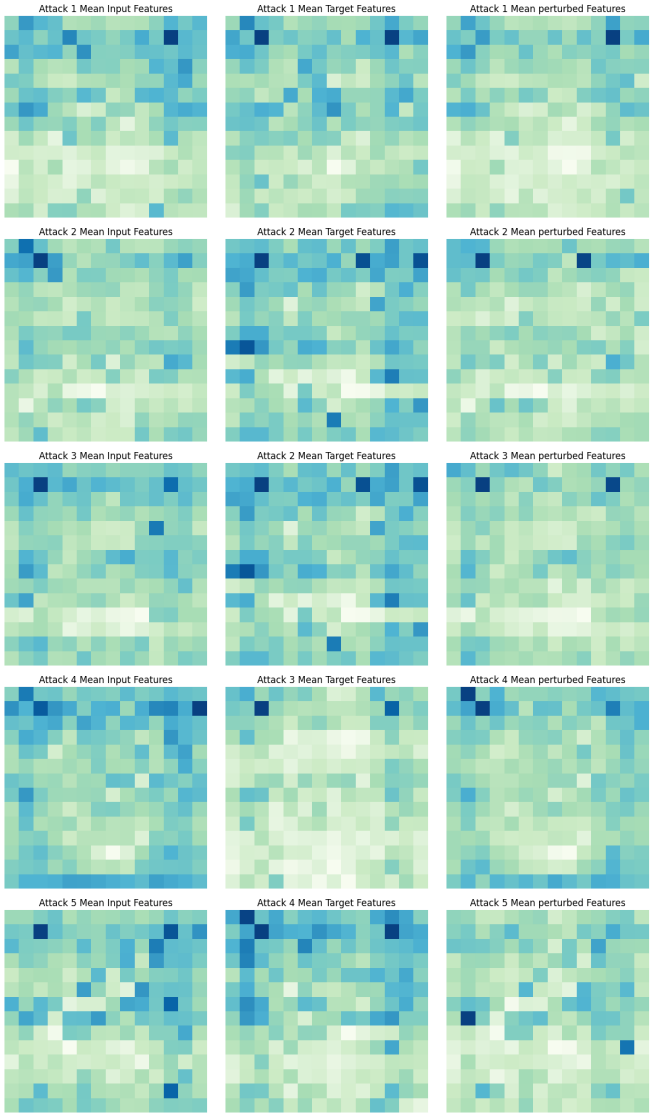

Fig. 5. The mean feature maps of the input, target, and perturbed images for Attacks 1-5.

V. Conclusion

The study found that vision transformers are vulnerable to FMViTA, which can significantly degrade their output accuracy. Therefore, developing more robust ViTs is crucial for ensuring the accuracy and reliability of ViT models. This highlights the importance of developing robust defense mechanisms to mitigate the impact of such attacks on vision transformers. However, investigating ViTs' robustness by generating adversarial attacks is a key to crafting robust defense mechanisms.

For future work, we will develop this direction of generating adversarial perturbations to be more robust to defense mechanisms. Furthermore, we will continue improving the FMViTA's performance in comparison with other state-of-the-art attacks. In addition, transferable improvements in the attack to be run in other ViT models are essential. Further, we can extend the FMViTA method to create other types of adversarial attacks, such as backdoor and poisoning attacks.

Acknowledgment

This article is derived from a research grant funded by the Research, Development, and Innovation Authority (RDIA), Kingdom of Saudi Arabia, with grant number 12615-iu-2023-IU-R-2-1-EI-.

References

[1] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *arXiv e-prints*, p. arXiv:1511.08458, Nov. 2015.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv e-prints*, p. arXiv:2010.11929, Oct. 2020.

[3] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences*, vol. 13, no. 9, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5521

[4] G. Kim, J. Kim, and J. Lee, "Exploring adversarial robustness of vision transformers in the spectral perspective," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, jan 2024, pp. 3964–3973. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00393

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv e-prints*, p. arXiv:1412.6572, Dec. 2014.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Machine Learning at Scale," *arXiv e-prints*, p. arXiv:1611.01236, Nov. 2016.

[7] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0041555364901375

[8] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting Adversarial Attacks with Momentum," *arXiv e-prints*, p. arXiv:1710.06081, Oct. 2017.

[9] X. Wang, J. Lin, H. Hu, J. Wang, and K. He, "Boosting adversarial transferability through enhanced momentum," in *British Machine Vision Conference*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232290454

[10] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1924–1933.

[11] R. Duan, Y. Chen, D. Niu, Y. Yang, A. K. Qin, and Y. He, "Advdrop: Adversarial attack to dnns by dropping information," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7486–7495.

[12] N. Park and S. Kim, "How Do Vision Transformers Work?" *arXiv e-prints*, p. arXiv:2202.06709, Feb. 2022.

[13] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8681–8691.

[14] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, *A fourier perspective on model robustness in computer vision*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[15] C. Guo, J. S. Frank, and K. Q. Weinberger, "Low Frequency Adversarial Perturbation," *arXiv e-prints*, p. arXiv:1809.08758, Sep. 2018.

[16] Y. Sharma, G. W. Ding, and M. A. Brubaker, "On the effectiveness of low frequency perturbations," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, ser. IJCAI'19. AAAI Press, 2019, p. 3389–3396.

[17] Z. Jin, J. Zhang, Z. Zhu, X. Wang, Y. Huang, and H. Chen, "Leveraging Information Consistency in Frequency and Spatial Domain for Adversarial Attacks," *arXiv e-prints*, p. arXiv:2408.12670, Aug. 2024.

[18] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv e-prints*, p. arXiv:1312.6199, Dec. 2013.

[19] M. Altoub, F. AlQurashi, T. Yigitcanlar, J. M. Corchado, and R. Mehmood, "An ontological knowledge base of poisoning attacks on deep neural networks," *Applied Sciences*, vol. 12, no. 21, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/21/11053

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[21] A. Aldahdooh, W. Hamidouche, and O. Deforges, "Reveal of Vision Transformers Robustness against Adversarial Attacks," *arXiv e-prints*, p. arXiv:2106.03734, Jun. 2021.

[22] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh, "On the Adversarial Robustness of Vision Transformers," *arXiv e-prints*, p. arXiv:2103.15670, Mar. 2021.

[23] S. Paul and P. Chen, "Vision transformers are robust learners," *CoRR*, vol. abs/2105.07581, 2021. [Online]. Available: https://arxiv.org/abs/2105.07581

[24] A. Joshi, G. Jagatap, and C. Hegde, "Adversarial token attacks on vision transformers," *CoRR*, vol. abs/2110.04337, 2021. [Online]. Available: https://arxiv.org/abs/2110.04337

[25] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[27] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K Pretraining for the Masses," *arXiv e-prints*, p. arXiv:2104.10972, Apr. 2021.

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv e-prints*, p. arXiv:1412.6980, Dec. 2014.

[29] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462 – 466, 1952. [Online]. Available: https://doi.org/10.1214/aoms/1177729392

[30] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv e-prints*, p. arXiv:1711.05101, Nov. 2017.

[31] J. Yang, K. Z. Luo, J. Li, C. Deng, L. Guibas, D. Krishnan, K. Q. Weinberger, Y. Tian, and Y. Wang, "Denoising Vision Transformers," *arXiv e-prints*, p. arXiv:2401.02957, Jan. 2024.